2.2 Mb of contiguous nucleotide sequence from chromosome III of *C. elegans*

R. Wilson*, R. Ainscough†, K. Anderson*, C. Baynes†, M. Berks†, J. Bonfield†, J. Burton†, M. Connell*, T. Copsey†, J. Cooper*, A. Coulson†, M. Craxton†, S. Dear†, Z. Du*, R. Durbin†, A. Favello*, A. Fraser†, L. Fulton*, A. Gardner†, P. Green*, T. Hawkins†, L. Hillier*, M. Jier*, L. Johnston*, M. Jones†, J. Kershaw†, J. Kirsten*, N. Laisster†, P. Latreille*, J. Lightning†, C. Lloyd†, B. Mortimore†, M. O'Callaghan†, J. Parsons*, C. Percy†, L. Rifken*, A. Roopra*, D. Saunders†, R. Shownkeen†, M. Sims†, N. Smaldon†, A. Smith†, M. Smith†, E. Sonnhammer†, R. Staden†, J. Sulston†, J. Thierry-Mieg‡, K. Thomas†, M. Vaudin*, K. Vaughan*, R. Waterston*, A. Watson†, L. Weinstock*, J. Wilkinson-Sproat† & P. Wohldman*

* Department of Genetics and Genome Sequencing Center, Washington University School of Medicine, St Louis, Missouri 63110, USA
† MRC Laboratory of Molecular Biology and Sanger Center, Hinxton Hall, Cambridge CB10 1RQ, UK
‡ CNRS-CRBM et Physique-Mathematique, Montpellier 34033, France

As part of our effort to sequence the 100-megabase (Mb) genome of the nematode *Caenorhabditis elegans*, we have completed the nucleotide sequence of a contiguous 2,181,032 base pairs in the central gene cluster of chromosome III. Analysis of the finished sequence has indicated an average density of about one gene per five kilobases; comparison with the public sequence databases reveals similarities to previously known genes for about one gene in three. In addition, the genomic sequence contains several intriguing features, including putative gene duplications and a variety of other repeats with potential evolutionary implications.

THE free-living nematode *Caenorhabditis elegans* is an excellent model organism for the study of development and behaviour, and its small size and short life cycle greatly facilitate genetic analysis[1,2]. Because a nearly complete clonal physical reconstruction of the genome is available[3–5], and 1,200 genetic loci have been identified, the nematode is also a good candidate for complete DNA sequence analysis. The haploid genome of *C. elegans* contains approximately 100 Mb (megabases) distributed on six chromosomes[2]. Many of the genes required for normal development and behaviour in the nematode have extensive similarity to their mammalian counterparts. However, compared with mammalian genes, genes in *C. elegans* typically have smaller and fewer introns[2,6], thus simplifying the identification of previously uncharacterized genes. Many of these newly identified genes may in turn be used to probe for as-yet unidentified mammalian genes.

As previously reported[6], we have embarked on a collaborative project to sequence the entire genome of *C. elegans*. Here we report some of the results from the first three years of the pilot phase, a point at which each laboratory has completed one contiguous megabase of genomic sequence. The combined data represent a sequence spanning more than 2.1 Mb. Most of the sequence derives from cosmid clones that were mapped to chromosome III by restriction fingerprinting[3]. At this physical map location, there are two large cosmid contigs, each more than 1 Mb long and bridged by a yeast artificial chromosome (YAC) clone. A small cosmid contig of 92 kilobases (kb) lies near the centre of the YAC bridge. Our genomic sequence analysis of this large region of chromosome III has confirmed the high gene density that we found in the first three sequenced cosmids[6], and has resulted in the identification of many more genes and other interesting sequence features.

Sequencing strategy

At the start of the pilot phase, we experimented with a primer-directed or 'walking' strategy in which site-specific oligonucleotide primers were used to extend sequences sequentially from a limited number of starting points[7]. This was initially done using cosmid DNA as template for the sequencing reactions. However, because cosmid DNA proved difficult to purify in sufficient quantities and was troublesome because of the presence of repeated sequences, we used random phagemid and M13 subclones as templates for the walking strategy[8,9]. During this work, we found that a combination of small insert (1–2 kb) and large insert (6–9 kb) subclones provided representative coverage for a typical cosmid. Even with subclones, primer-directed sequencing was occasionally problematic because of the presence of repeated sequences. Thus, our strategy has evolved to the point where most sequence data come from the initial readings of 600–800 random subclones. As the use of automated fluorescent DNA sequencing machines for data collection provides high-throughput sample processing[10,11], this approach is efficient and cost-effective. This random or 'shotgun' phase not only provides much of the final sequence, but also maps the subclones needed for closing gaps and completing the complementary strand. At this point, a walking strategy can be successfully exploited for completion, as most repeated sequences in the cosmid insert can be identified and selectively avoided. Many of the details of our sequencing strategy have been reported elsewhere[6,12,23].

Sequences were considered to be finished when every base had been determined on both strands and all ambiguities had been resolved. At this point, the finished sequence was compared with the public sequence databases using the BLASTX program for protein similarities and BLASTN for nucleotide similarities[24]. When similarity searches were complete, likely genes were identified using GENEFINDER (P.G. and L.H., unpublished), and in some cases interactively edited using ACEDB (R.D. and J.T.-M., unpublished). Finished sequences were annotated with regard to likely genes, homologies, *trans*-spliced leaders[26], complementary DNA matches[27,28] and other features, such as structural RNA genes and transposons, and submitted to the GenBank and EMBL databases. Also, to present genomic sequence data in the context of the physical and genetic maps of the nematode, all sequences are deposited in the *C. elegans* database ACEDB, which is available to the research community.

Sequences

Table 1 gives the cosmid clones that were sequenced, along with their database accession numbers and lengths. Most of the redundant overlapping data have been omitted from the data-

TABLE 1 Sequences submitted to the GenBank and EMBL databases

Cosmid name	Locus name	Acc. no.	Length (bp)	Cosmid name	Locus name	Acc. no.	Length (bp)
ZK112	CELZK112	L14324	38,269	K02D10	CELK02D10	L14710	18,683
ZC97	CELZC97	L14714	4,166	F54F2	CELF54F2	L23645	39,573
ZK686	CELZK686	L17337	11,435	F44E2	CELF44E2	L23646	33,651
C08C3	CELC08C3	L15201	44,025	pAR3	CELPAR3	U00026	4,590
C27D11	CELC27D11	L23650	9,973	K01F9	CEK01F9	Z22175	19,834
ZK652	CELZK652	L14429	36,052	ZK637	CE1	Z11115	40,699
C02C2	CELC02C2	L23649	20,495	ZK638	CEZK638	Z12018	1,762
ZK688	CELZK688	L16621	36,977	ZK643	CEZK643	Z11126	39,534
C29E4	CELC29E4	L23651	40,050	R08D7	CER08D7	Z12017	27,368
F54H12	CELF54H12	L25599	19,168	F59B2	CEF59B2	Z11505	43,782
C06G4	CELC06G4	L25598	29,122	R107	CER107	Z14092	40,970
F44B9	CELF44B9	L23648	36,327	F02A9	CEF02A9	Z19555	26,242
K12H4	CELK12H4	L14331	38,582	ZK507	CEZK507	Z29116	13,501
K06H7	CELK06H7	L15314	22,073	ZK512	CEZK512	Z22177	36,997
C14B9	CELC14B9	L15188	43,492	F54G8	CEF54G8	Z19155	31,613
D2007	CELD2007	L16560	13,651	ZC84	CEZC84	Z19157	38,955
C50C3	CELC50C3	L14433	44,733	T23G5	CET23G5	Z19158	26,926
C30A5	CELC30A5	L10990	27,743	T02C1	CET02C1	Z19156	10,308
C02F5	CELC02F5	L14745	22,333	M01A8	CEM01A8	Z27081	19,001
F09G8	CELF09G8	L11247	41,449	K01B6	CEK01B6	Z22174	34,002
F10E9	CELF10E9	L10986	32,733	C40H1	CEC40H1	Z19154	27,271
ZC262	CELZC262	L23647	4,166	K04H4	CEK04H4	Z27078	33,930
R05D3	CELR05D3	L07144	38,810	C38C10	CEC38C10	Z19153	34,193
ZK353	CELZK353	L15313	24,916	T26G10	CET26G10	Z29115	30,251
ZK1236	CELZK1236	L13200	28,878	F54C8	CEF54C8	Z22178	23,000
C30C11	CELC30C11	L09634	30,865	B0464	CEB0464	Z19152	40,090
F42H10	CELF42H10	L08403	28,687	F55H2	CEF55H2	Z27080	22,950
C04D8	CELC04D8	L16687	10,433	ZK1098	CEZK1098	Z22176	37,310
ZC21	CELZC21	L16685	36,087	C48B4	CEC48B4	Z29117	35,000
K10C7 (C02D5)	CELC02D5	L16622	28,735	F58A4	CEF58A4	Z22179	38,000
C06E1/F43A9	CELC06E1	L16560	40,216	C15H7	CEC15H7	Z22173	28,000
C13G5	CELC13G5	L14730	10,883	C07A9/C40D10	CEC07A9	Z29094	66,004
F22B7	CELF22B7	L12018	40,222	T05G5	CET05G5	Z27079	36,180
B0523	CELB0523	L07143	14,334	R10E11	CER10E11	Z29095	32,254
B0303	CELB0303	M77697	41,071	ZK632	CEZK632	Z22181	36,000
ZK370	CELZK370	M98552	37,675	K11H3	CEK11H3	Z22180	33,000
pAR2	CELPAR2	U00025	12,721	ZK757	CEZK757	Z29121	31,000

base entries, although neighbouring sequences typically contain a small number of overlapping bases to facilitate joining or to keep a gene intact. The total non-overlapping sequence assembled from all of the clones in Table 1 spans 2.181 Mb. A map of this region showing the sites of predicted genes and previously known loci is presented in Fig. 1. Several additional cosmids which extend the sequence more than 2,000 kb to the left and 500 kb to the right are in various stages of completion.

The most striking result from genomic DNA sequencing is the continued high number of predicted genes in the region. In the 2.181 Mb of genomic sequence reported here, 483 putative genes were identified by similarity searches and GENEFINDER analysis. Start points for these candidate genes are indicated for both strands of the genomic sequence in Fig. 1. Generally, genes seem to be dispersed evenly throughout the region, with only a few examples of apparent clustering. One of the most gene-dense regions is contained in cosmids C30A5, C02F5 and F09G8 (Fig. 1, 492–553 kb). Here a 61-kb region contains 141 exons (in 24 putative genes), which account for 47% of the sequence. When the introns are also considered, more than 80% of the bases in this region are within predicted genes. Also in a 69.5-kb stretch of genomic DNA (F55H2, ZK1098; Fig. 1, 1,770–1,840 kb), 109 exons (in 19 putative genes) are predicted, with 40% of this region representing coding sequence and a total of 65% contained in genes. The longest stretch of genomic sequence that does not contain a likely gene is ~25 kb (pAR3, K01F9; Fig. 1, 1,145–1,170 kb). By comparison, the entire 2.181-Mb region is 29% coding sequence, with a total of 48% representing putative exons and introns. In our analysis of the first 0.1% of the genome, genes were found every 3–4 kb on average[6]. With more than 2% of the genomic DNA sequence now completed, the density is one gene every 4.5 kb. Because the region is expected to be relatively gene-rich[2], we cannot use this density to extrapolate the total gene number. However, an estimate independent of gene density can be made using the number of tagged cDNAs which hit candidate genes in the sequence[27]. Because 125 of the 4,615 *C. elegans* cDNA tags match predicted genes in the sequence, we can now estimate that the genome contains about 17,800 genes ($483 \times 4615/125$) for an average density of one gene every 5.6 kb.

There are no clear examples of genes that overlap, either on the same or on complementary strands. Several cases of orphan open reading frames that have high GENEFINDER scores and that overlap or are contained within candidate genes were observed, although there are no data to indicate that any of these is expressed. When the translated sequence is used in all six reading frames to search the public databases using the program BLAST, approximately one-third of the genes find significant matches to proteins from organisms other than *C. elegans*, with several very highly conserved genes indicated (Table 2). As can be seen in Table 2, a wide variety of different functions are represented, and with the exception of a homeobox cluster (see below), there is no obvious clustering of genes with related function. Most of the matches represent cross-phylum matches, for very little sequence from other nematodes is present in the databases. The fraction of genes with cross-phylum matches will increase as more genes from other organisms are entered into the database, but our previous analysis of these ancient conserved regions indicates that it is unlikely to rise above 40%, as most ancient conserved regions are already represented in the databases[29]. Although the data are inconclusive, preliminary evidence suggests that the fraction of matches to vertebrate proteins will be similar to this for any non-vertebrate genome.

Known genes

Several interesting genes had been positioned in this region before our sequence analysis. The first 120 kb of the sequence reported here is part of the HOM homeobox gene complex which extends an additional 150 kb to the left[30–33]. The region is centred around the *Antennapedia*-like gene *mab-5*, which is responsible

TABLE 2 Gene candidates showing significant similarity to existing protein sequences

Gene name	Closest DB hit, Acc. no.	Closest block	Description
ZK112.2	TVMSBF, A40951	—	Kinase-related transforming protein
ZK112.7	A41087	BL00232	Tumour suppressor
ZK686.2	DB73_DROME, P26802	BL00039	ATP-dependent RNA helicase
ZK686.8	A24148	—	*N*-acetyllactosamine synthetase
C08C3.1	HM11_CAEEL, P17486	BL00027	*C. elegans* Hox protein egl-5 (ceh-11) gene
C08C3.3	HMMA_CAEEL, P10038	BL00027	*C. elegans* Hox protein mab-5
ZK652.4	R5RT35, A34571	BL00579	60S ribosomal protein L35
ZK652.5	A34218	BL00027	Distal-less homeotic protein
ZK652.6	S29962	—	ref(2)P protein, Zn-finger region
ZK652.8	C35815	—	Myosin heavy chain-3
C02C2.1	TYRO_STRGA, P06845	BL00497	Tyrosinase
C02C2.3	ACHG_RAT, P18916	BL00236	Acetylcholine receptor
C02C2.4	S27951	—	Sodium/phosphate transport protein
ZK688.8	A24148	—	*N*-acetyllactosamine synthase
C29E4.3	RNA1_YEAST, P11745	—	RNA production/processing
C29E4.7	S16267	—	Auxin-induced protein
C29E4.8	JS0422, JS0422	BL00113	Adenylate kinase
F54H12.1	ACON_YEAST, P19414	BL00450	Aconitate hydratase
F54H12.6	EF1B_BOMMO, P29522	—	Elongation factor 1β
C06G4.2	CAP3_RAT, P16259	—	Calpain
C06G4.5	SSR3_MOUSE, P30935	BL00237	Somatostatin receptor
F44B9.1	ACPH_RAT, P13676	BL00708	Acylamino-acid-releasing enzyme
F44B9.8	A45253	—	Replication factor C
F44B9.9	PARB12, S11060	BL00125	Protein phosphatase
K12H4.1	JQ1397	—	*Drosophila melanogaster* Prospero
K12H4.4	SPC2_CHICK, P28687	—	Signal peptidase
K12H4.8	DEAD_ECOLI, P23304	BL00039	ATP-dependent RNA helicase dead
K06H7.1	S22127	BL00107	Protein kinase
K06H7.3	IPPI_YEAST, P15496	—	Isopentenyl-diphosphate δ-isomerase
K06H7.4	S24168	—	Sec7
K06H7.8	YCK1_YEAST, P23291	—	Casein kinase I
C14B9.1	CRAB_HUMAN, P02511	—	α-B-crystallin
C14B9.2	ER72_MOUSE, P08003	BL00194	Deoxycytidine kinase
C14B9.4	S22127, S22127	BL00098	Protein kinase
C14B9.7	R5RT21, A33295	—	Ribosomal protein L21
C14B9.8	S24109	—	Phosphorylase kinase
D2007.5	KERB_AVIER, P00535	—	ERB-B tyrosine kinase
C50C3.3	SPCN_CHICK, P07751	BL00545	Spectrin α-chain
C50C3.5	TPC1_BALNU, P21797	BL00018	Calmodulin
C50C3.7	OCRL_HUMAN, Q01986	—	Inositol polyphosphate-5-phosphatase
C50C3.11	JH0565	—	Calcium channel α-2b chain
C30A5.1	GRR1_YEAST, P24814	—	GRR1 protein (same as C02F5.7)
C30A5.3	S30854	BL00125	Phosphoprotein phosphatase
C30A5.4	SYB_DROME, P18489	BL00417	Synaptobrevin
C30A5.6	UN86_CAEEL, P13528	BL00035	Unc-86 alternate protein
C30A5.7	UN86_CAEEL, P13528	BL00035	Unc-86 protein
C02F5.3	JC1349	—	GTP-binding protein
C02F5.7	GRR1_YEAST, P24814	—	Glucose-induced repressor (GRR1)
C02F5.9	PRC5_HUMAN, P20618	BL00631	Proteasome component C5
F09G8.3	RS9_BACST, P07842	BL00360	Ribosomal protein S9
F09G8.6	A37122	—	Cuticle collagen
ZC262.3	A43425	—	N-CAM Ig domain
ZC262.5	ATPE_BOVIN, P05632	—	ATP synthase ε-chain
R05D3.1	TOPB_HUMAN, Q02880	BL00177	DNA topoisomerase II homologue
R05D3.3	ZG44_XENLA, P18721	—	Gastrula zinc-finger protein
R05D3.6	ATPE_BOVIN, P05632	—	ATP synthase ε-chain
R05D3.7	KINH_LOLPE, P21613	BL00411	Kinesin heavy chain
ZK353.6	AMPA_RICPR, P27888	BL00631	Aminopeptidase
ZK1236.1	LEPA_ECOLI, P07682	BL00301	LepA
ZK1236.2	NUCL_RAT, P13383	—	Nucleolin
C30C11.2	DXA2_MOUSE, P14685	—	Diphenol oxidase
C30C11.4	S30788	BL00297	HSP Msi3p
F42H10.4	GYRTI, A03270	—	Cysteine-rich intestinal protein
C04D8.1	SPCB_DROME, Q00963	—	Spectrin β-chain
ZC21.2	JH0588	—	Trp protein
ZC21.3	S14548	—	Dual bar protein
ZC21.4	S29956	—	Breakpoint cluster region (Bcr) protein
C02D5.1	ACDL_RAT, P15650	BL00072	Acyl-CoA dehydrogenase
C06E1.10	S22609, S22609	BL00690	Hypothetical protein
C06E1.4	B40171	—	Glutamate receptor
C06E1.8	B60191	—	Zn-finger
C06E1.9	A31922	—	ATP-dependent RNA helicase
C13G5.1	F34510	BL00027	Engrailed homeotic protein
F22B7.4	S18345	—	Environmental stress protein
F22B7.5	DNAJ_ECOLI, P08622	BL00636	DnaJ
F22B7.6	MUCB_ECOLI, P07375	—	Mucb protein
F22B7.7	S09048	—	Potassium channel protein Hak-6
B0523.1	S00904	BL00239	Tyrosine kinase
B0523.5	GELS_MOUSE, P13020	—	Gelsolin (flightless-1)
B0303.1	CYYA_YEAST, P08678	—	Adenylate cyclase
B0303.2	PNMT_BOVIN, P10938	—	Phenylethanolamine-*N*-methyltransferase
B0303.3	THIL_RAT, P17764	BL00098	Acetyl-CoA acetyltransferase
B0303.5	YT31_CAEEL, P03934	—	Tc3
B0303.7	NCF2_HUMAN, P19878	—	SH3 domain
B0303.9	SLP1_YEAST, P20795	—	SLP-1
ZK370.3	TALI_MOUSE, P26039	—	Talin
ZK370.4	KAPR_DICDI, P05987	—	Cyclic AMP-dependent protein kinase
ZK370.5	BCKD_RAT, Q00972	—	3-Methyl-2-oxobutanoate dehydrogenase
K02D10.1	S16088, S16088	—	4-Nitrophenylphosphatase
K02D10.5	S07258, S07258	—	*Escherichia coli* plasmid *RK2* gene for P116
F54F2.1	ITAP_DROME, P12080	BL00242	Vitronectin receptor α-subunit

TABLE 2—Continued

Gene name	Closest DB hit, Acc. no.	Closest block	Description
F54F2.2	A44067	—	109K basic protein H
F44E2.1	S08405	—	Protease
F44E2.3	S28589	—	DnaJ
F44E2.4	S03430	—	LDL receptor
F44E2.6	PILB_NEIGO, P14930	—	PILB protein
F44E2.7	CALD_CHICK, P12957	—	Caldesmon
ZK637.1	STP1_ARATH, P23586	—	Sugar transporter
ZK637.10	GSHR_HUMAN, P00390	BL00076	Glutathione reductase
ZK637.11	CC25_SCHPO, P06652	—	CDC25
ZK637.13	GLBH_TRICO, P27613	—	Globin
ZK637.14	PIC0_HSV11, P08393	BL00518	Transactivator ICP0 (motif 1)
ZK637.5	ARSA_ECOLI, P08690	—	ArsA
ZK637.8	VPP1_RAT, P25286	—	Proton pump
ZK643.2	DCTD_BPT2, P00814	—	DCMP deaminase
ZK643.3	CALR_PIG, P25117	BL00649	G-protein-coupled receptor
R08D7.5	CATR_CHLRE, P05434	BL00018	Calcium-binding protein
R08D7.6	CNAG_BOVIN, P14099	BL00126	Cyclic GMP phosphodiesterase
F59B2.3	NAGA_ECOLI, P15300	—	*N*-acetyl-glucosamine-6-phosphate deacetylase
F59B2.7	RAB6_HUMAN, P20340	—	Rab6 (Ras protein)
R107.7	GTP_CAEEL, P10299	—	Glutathione S-transferase P subunit
LIN12A.cds	LI12_CAEEL, P14585	BL00022	Lin-12/Notch, EGF and ankyrin repeats
F02A9.5	PCCB_RAT, P07633	—	Propionyl-CoA carboxylase
GLP1A.cds	GLP1_CAEEL, P13508	BL00022	Lin-12/Notch, EGF and ankyrin repeats
ZK507.1	HR25_YEAST, P29295	BL00107	HRR25 protein kinase
ZK507.6	CG2A_PATVU, P24861	BL00292	G2/M cyclin A
ZK512.2	SPB4_YEAST, P25808	BL00039	RNA helicase
ZK512.4	SRP9_CANFA, P21262	—	Signal recognition particle 9K protein
F54G8.2	KDGL_PIG, P20192	BL00479	Diacylglycerol kinase
F54G8.3	ITA3_CRISP, P17852	BL00242	Integrin α-chain
F54G8.4	KRET_HUMAN, P07949	BL00518	Ret zinc-finger region
ZC84.1	LACI_RABIT, P19761	BL00280	Serine protease inhibitor
ZC84.2	CNGC_RAT, Q00195	—	Cyclic nucleotide gated olfactory channel
ZC84.4	GASR_RAT, P30553	BL00237	G-protein-coupled receptor
T23G5.1	RIR1_HUMAN, P23921	BL0089	Ribonucleoside-disphosphate reductase lg chain
T23G5.2	SC14_KLULA, P24859	—	SEC14 (yeast)
T23G5.5	NTTN_HUMAN, P23975	BL00610	Neurotransmitter transporter
T02C1.1	RA18_YEAST, P10862	BL00518	RAD-18 DNA-binding protein
M01A8.4	BIK1_YEAST, P11709	—	Nuclear fusion protein
C40H1.1	S24577	—	Ovarian protein (*D. melanogaster*)
C40H1.4	YCS4_YEAST, P25358	BL00030	Yeast hypothetical protein
K04H4.1	CA14_CAEEL, P17139	—	Collagen
C38C10.1	NK3R_RAT, P16177	BL00237	G-protein-coupled receptor
C38C10.5	RGR1_YEAST, P19263	—	Glucose repression regulatory protein RGR1
T26G10.1	DEAD_ECOLI, P23304	BL00039	RNA helicase
T26G10.3	RS24_HUMAN, P16632	BL00529	Ribosomal protein S24
F54C8.1	HCDH_PIG, P00348	BL00067	3-hydroxyacyl-CoA dehydrogenase
F54C8.2	H31_SCHPO, P09988	BL00322	Histone H3
F54C8.4	Y19K_NPVAC, P24656	—	ACMNPV hypothetical protein
F54C8.5	RAS_LENED, P28775	—	Ras family
B0464.1	SYD2_HUMAN, P14868	BL00179	Aspartyl-tRNA synthetase
B0464.5	KCLK_MOUSE, P22518	BL00107	Serine/threonine kinase
F55H2.1	SODC_BOVIN, P00442	BL00087	Superoxide dismutase
F55H2.2	MTPG_SULAC, P22721	—	Membrane-associated ATPase γ-chain
F55H2.5	C561_BOVIN, P10897	—	Cytochrome b_{561}
ZK1098.10	TPMX_RAT, P18342	—	Coiled-coil protein
ZK1098.4	GCN3_YEAST, P14741	—	GCN3 (yeast transcription activator)
C48B4.1	CA01_RAT, P07872	BL00072	Acyl-CoA oxidase I
C48B4.2	RHOM_DROME, P20350	—	Rhomboid (*D. melanogaster*)
C48B4.4	NODI_RHILO, P23703	BL00211	ATP-binding transport protein
C48B4.5	LIVG_SALTY, P30293	BL00211	ATP-binding transport protein
F58A4.10	UBC7_WHEAT, P25868	BL00183	Ubiquitin-conjugating enzyme
F58A4.3	H3_VOLCA, P08437	BL00322	Histone H3
F58A4.4	PRI1_MOUSE, P20664	—	DNA primase 49K subunit
F58A4.5	RTJK_DROME, P21328	—	Mobile element Jockey-rev. transcriptase
F58A4.7	A36394	BL00038	Transcription factor AP-4
F58A4.8	TBG_XENLA, P23330	BL00227	γ-Tubulin
F58A4.9	RPC9_YEAST, P28000	—	RNA Pol I/III 16K polypeptide
C15H7.2	KFPS_DROME, P18106	BL00790	Tyrosine kinase
C15H7.3	TCPT_HUMAN, P17706	—	Protein tyrosine phosphatase
C07A9.2	G10_XENLA, P12805	—	G10 protein
C07A9.3	KRAC_HUMAN, P31749	BL00107	Ser/Thr kinase
C07A9.4	S20969	BL00470	Na/Ca, K antiporter
C07A9.5	SPCA_DROME, P13395	BL00018	Spectrin α-chain
C07A9.6	UDP2_RAT, P09875	BL00375	UDP-glucuronosyltransferase
T05G5.3	CC2_HUMAN, P06493	BL00107	CDC2 kinase
T05G5.5	S27735	—	Hypothetical protein A (*Thermus aquaticus*)
T05G5.6	ECHM_RAT, P14604	BL00166	Enoyl-CoA hydratase
T05G5.10	IF5A_HUMAN, P10159	—	Initiation factor 5A
R10E11.1	FSH_DROME, P13709	BL00633	Bromodomain
R10E11.2	VATL_DROME, P23380	BL00605	Vacuolar ATP synthase subunit
R10E11.3	UBP2_YEAST, Q01476	—	Ubiquitin-specific processing protease
R10E11.4	NALS_MOUSE, P15535	—	Galactosyltransferase
ZK632.1	MCM3_YEAST, P24279	—	Mcm 2/3
ZK632.3	S26727	—	Hypothetical protein 186 (*Thermoplasma acidophilum*)
ZK632.4	MANA_ECOLI, P00946	—	Mannose 6-phosphate isomerase
ZK632.6	CALX_HUMAN, P27824	BL00803	Calnexin
ZK632.8	ARF2_YEAST, P19146	—	ADP-ribosylation factor
K11H3.1	GPDA_DROVI, P07735	—	Glycerol-3-phosphate dehydrogenase
K11H3.3	UCP_HUMAN, P25874	BL00215	Mitochondrial carrier family
ZK757.2	TPCL_HUMAN, P28562	—	Protein tyrosine phosphatase

for pattern formation in a posterior body region[31]. GENEFINDER predicted a gene candidate with sequence identity to the *mab-5* cDNA sequence (GenBank M22751) on the complementary strand of the cosmid clone C08C3 (annotated as C08C3.3). Approximately 30 kb to the right of *mab-5*, the *abdominal-B*-like gene *egl-5*(*ceh-11*)[31,32], which is required for normal development of several cell types in the tail region, was located (C08C3.1). Interestingly, the *mab-5* and *egl-5* genes are encoded in opposite orientation. A third homeobox gene, *ceh-23*, lies 23 kb to the right of *egl-5* (ZK652.5). The *ceh-23* gene, which is similar to the *Drosophila* genes *Distal-less* and *empty spiracles*, was located by identity to a cDNA clone[32]. Two additional homeobox genes, *lin-39*(*ceh-15*) and *ceh-13*, lie approximately 200 kb to the left of *mab-5*. Preliminary sequence data indicate that the order of these two genes relative to *mab-5* is *lin-39, ceh-13*, and not *ceh-13, lin-39* as originally reported[32,33]. An unrelated gene, *egl-45*, with no similarity to any known *Drosophila* genes, maps between the *egl-5* and *ceh-23* genes (tentatively correlated to gene candidate *C27D11.1*), and several other putative genes are contained within the HOM region.

Other genes previously mapped to the region were correlated to distinct loci (shown in parentheses) by genomic DNA sequencing. These include *egl-45* (tentatively C27D11.1), *lin-36* (F44B9.6), *unc-36* (C50C3.11), *unc-86* (C30A5.7), *mig-10* (tentatively F10E9.6), *unc-116* (R05D3.7), *ceh-16* (C13G5.1), *dpy-19* (F22B7.10), *sup-5* (B0523.5), *unc-32* (ZK637.10), *lin-9* (ZK637.7), *gst-1* (R107.7), *lin-12* (LIN12A in cosmid R107), *glp-1* (GLP1A in cosmid F02A9), *emb-9* (K04H4.1), *tbg-1* (F58A4.8) and *ncc-1* (T05G5.3). The *unc-86* and *emb-9* genes and part of the *ceh-16* gene has been sequenced previously[34–36]. In the cases of *egl-45* (M. Basson and H. R. Horvitz, personal communication), *lin-36* (J. Thomas and H. R. Horvitz, personal communication), *unc-36* (L. Loebel and H. R. Horvitz, personal communication), *unc-116* (ref. 37), *lin-9* (G. Beitel and H. R. Horvitz, personal communication), *gst-1* (ref. 38), *lin-12* (ref. 39) and *glp-1* (ref. 40), cDNA sequences provided by researchers within the *C. elegans* community enabled gene assignments to be made. For *tbg-1* and *ncc-1*, cDNAs from the consortium tag-sequencing project[27] allowed assignment to genomic loci. For *mig-10, dpy-19* and *unc-32*, transgenic rescue experiments with mutant phenotypes localized the genes to a particular restriction fragment (J. Manser; S. Colloms; D. Thierry-Mieg, personal communications). In some cases, availability of the genomic sequence facilitated this type of analysis.

A few kilobases of genomic DNA sequence which included *sup-5* had been reported previously[41]. However, additional genomic sequencing revealed that the *sup-5* locus, a gene for transfer RNA^{Trp}, lies within an intron in the same transcriptional orientation as a homologue of *Drosophila melanogaster flightless I* (ref. 42) (B0303.1). As the *sup-5* mutation has been shown to suppress specific alleles of many unrelated genes[43], it is known to be a functional tRNA gene. Further, the *C. elegans fl I* homologue is expressed and spliced as predicted (ref. 42, and R. Wilson, unpublished data).

tRNA genes

The haploid genome of *C. elegans* has been estimated to contain about 300 tRNA genes[44]. Thus, we would expect to find an average of three tRNA genes per megabase of genomic sequence. However, using tRNAscan[45], in the 2.181 Mb of the sequence reported here, at least 14 tRNA genes were identified. These are indicated in Fig. 1. Strikingly, two cosmid clones, C14B9 and F22B7, contain seven of these tRNA genes. Like the *sup-5* tRNA gene, a $tRNA^{Ser}$ and two $tRNA^{Phe}$ genes lie within introns of likely genes.

Repeats

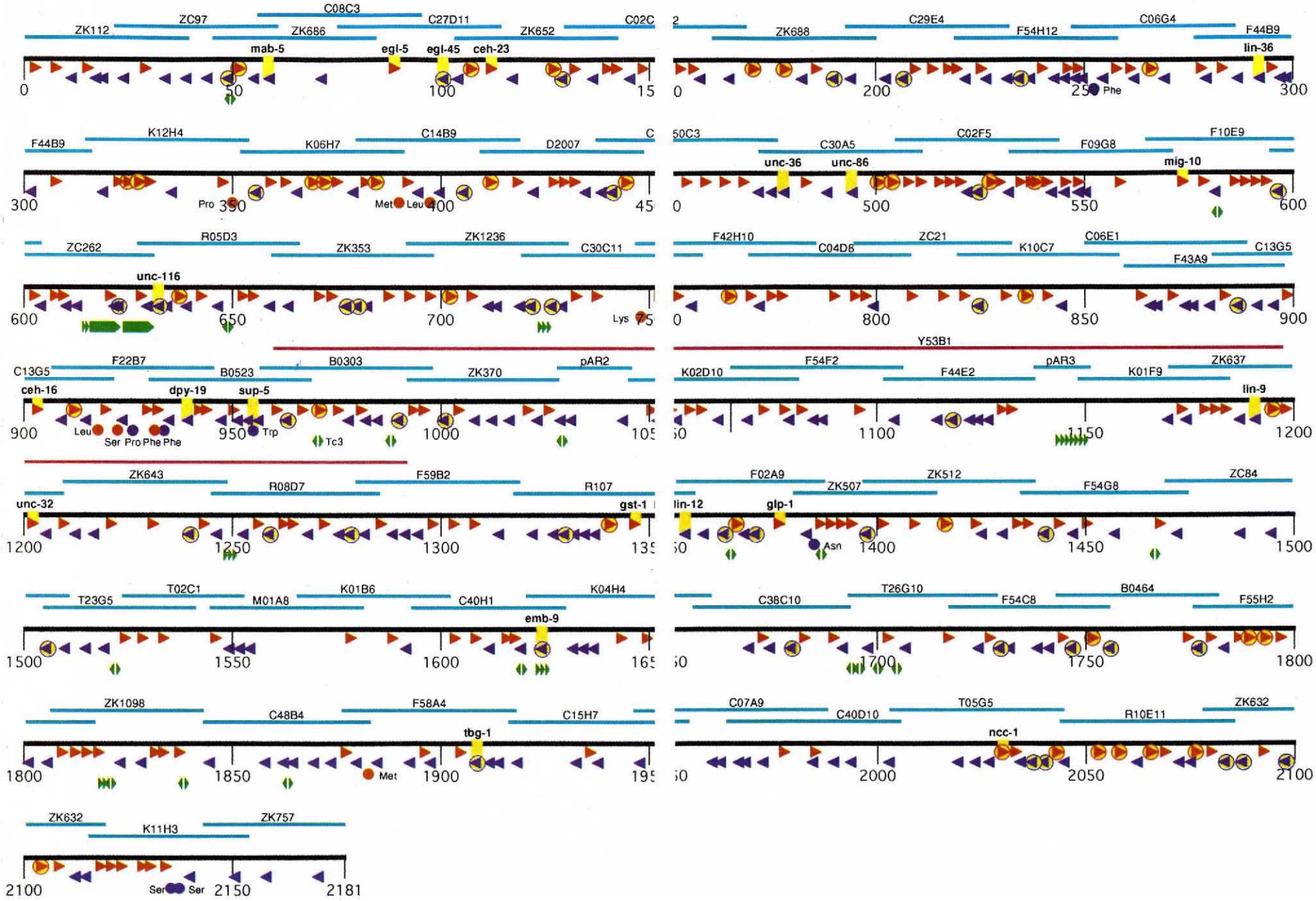
Several types of repeated sequence are present in the 2.181-Mb region. Figure 1 indicates the locations of the larger and more complex repeats. Detailed analysis of three major types of repeated sequences (inverted, tandem and interspersed) reveals several interesting features.

Inverted repeats, in which a segment of genomic sequence lies within a few to several hundred bases of an inverted copy of itself, are the most common type of repeat that we have found. Considering only those inverted repeats of up to 1 kb end to end, with at least 70% identity, on average an inverted repeat is found every 5.5 kb. Most of these are quite small, with an average segment length of 70 base pairs (bp) and an average loop size of 164 bp. A relatively high proportion of these repeats (43%) occur in introns, which represent only 20% of the genome. Most inverted repeats fall into families and may be remnants of mobile elements. In particular, there were examples of inverted repeat elements from known transposons Tc*3* (ref. 46) and Tc*4* (ref. 47).

Tandem repeats, in which a segment of genomic sequence lies adjacent to one or more copies of itself, occur on average every 10 kb. As with inverted repeats, most of these are small, with an average segment length of 17 bp and an average copy number of 14. Interestingly, only 17% of tandem repeats were found in introns, whereas 63% occurred between genes. The most common category of tandem repeats were triplets, some of which were found in predicted exons. One of the most complex tandem repeats found was a 95-bp sequence which was repeated more than 30 times in the clone pAR3 (Fig. 1; 1,144–1,150 kb). This region, which was missing from the cosmid map, had to be recovered from the YAC clone Y53B1 by targeted gap rescue in yeast. Also a large (7.9 kb) tandem repeat, flanked by more complex short repeats, was found in cosmid ZC262 (Fig. 1; 595–633 kb). Sequence assembly for these repeat regions was

FIG. 1 The region of chromosome III described here. Each line represents 300 kb, with cosmid clones indicated by blue bars; a magenta bar indicates the YAC clone Y53B1. Red and blue arrowheads indicate the approximate starting points of gene candidates for both strands. Circled arrowheads indicate cDNA hits. Red or blue circles indicate the positions of tRNA genes. Green arrowheads indicate the position and type of major repeats. Genetic markers previously mapped to the region and assigned to genomic loci after DNA sequencing are indicated by vertical yellow bars. ►

METHODS. Subclone libraries were prepared and sequencing reactions performed as previously described[6,12,14]. For data collection, automated fluorescent DNA sequencers were used. Both the Applied Biosystems 373A and Pharmacia ALF instruments were initially tested, although the 373A was preferred for the shotgun phase because of its greater sample capacity. Both laboratories currently complete two daily runs of each sequencer with 36 lane gels. Thus, with only four of these sequencing instruments, typically more than 1,700 samples are processed per week. After subtracting for failures and contaminating vector sequences, this is sufficient for yearly production of more than 20 Mb of raw sequence data. Two additional sequencing instruments are available for directed sequencing to close gaps and complete the complementary strand. At the conclusion of each sequencer run, the ABI trace files were transferred from the host Macintosh to UNIX workstations for all further work. During the shotgun phase all processing on the UNIX systems is fully automated using a single script. This includes reformatting the trace files using MAKESCF (S.D., unpublished), selection of the quality data from each reading using AUTOTED (L.H., unpublished), clipping off cosmid and sequencing vector sequences using VEP (R.S., unpublished), and assembly using BAP[13]. After the shotgun phase, XBAP[13], which includes the oligo selection engine of OSP[22], is used to edit and finish each project. GENEFINDER and COP (S.D., unpublished) were used to check the finished sequence for errors. GENEFINDER is useful for identifying insertions and deletions in regions containing predicted genes, and COP, which compares the final sequence back to the raw data from which it was produced, is useful for identifying editing errors. We and others have previously described an evaluation of the accuracy of raw data from automated fluorescent DNA sequencers[6,25].


C08C3
ZC97
C27D11
C02C
ZK112
ZK686
ZK652
mab-5
egl-5
egl-45
ceh-23
0
50
100
15
2
C29E4
C06G4
ZK688
F54H12
F44B9
lin-36
0
200
250
Phe
300
F44B9
K12H4
K06H7
C14B9
D2007
C
300
Pro
350
Met
Leu
400
45
50C3
C02F5
F10E9
C30A5
F09G8
unc-36
unc-86
mig-10
0
500
550
600
ZC262
R05D3
ZK353
ZK1236
C30C11
unc-116
600
650
700
Lys
75
F42H10
C04D8
ZC21
K10C7
C06E1
C13G5
F43A9
0
800
850
900
Y53B1
C13G5
F22B7
B0523
B0303
ZK370
pAR2
ceh-16
dpy-19
sup-5
900
Leu
Ser
Pro
Phe
Phe
950
Trp
Tc3
1000
105
K02D10
F54F2
F44E2
pAR3
K01F9
ZK637
lin-9
50
1100
1150
1200
ZK643
R08D7
F59B2
R107
unc-32
gst-1
1200
1250
1300
135
F02A9
ZK507
ZK512
F54G8
ZC84
lin-12
glp-1
50
Asn
1400
1450
1500
T23G5
T02C1
M01A8
K01B6
C40H1
K04H4
emb-9
1500
1550
1600
165
C38C10
T26G10
F54C8
B0464
F55H2
50
1700
1750
1800
ZK1098
C48B4
F58A4
C15H7
tbg-1
1800
1850
Met
1900
195
C07A9
C40D10
T05G5
R10E11
ZK632
ncc-1
50
2000
2050
2100
ZK632
K11H3
ZK757
2100
Ser
Ser
2150
2181

accomplished with very stringent parameters and, in some cases, aided by map information.

There are several examples of short repetitive sequences that are scattered throughout the genome, including the 94-bp consensus of the repeat element from most common inverted repeat families, previously observed in *lin-12* and *glp-1* introns[39,40]. These elements are widely dispersed, with an average of 14 copies every 100 kb at varying levels of conservation. Often, the repeat elements of inverted repeat families are also found in singleton or tandem arrangements. For example, a tandem pair of degenerate elements (69% identical over 77 bp) flanks part of the predicted gene *F58A4.2*. As discussed below, this seems to have been duplicated from a region 200 kb away.

Gene duplications

In addition to the short interspersed repeats discussed above, there are sequences that are repeated tens or hundreds of kilobases apart, with up to 98% similarity. In some cases, these apparent duplications have a complex structure wherein several segments from one region are repeated in a second location, but with different spacings and orientations. Many of these long-range repeats involve coding regions. In particular, there seems to be a recent gene duplication involving *F22B7.5* and *C38C10.4*, which are approximately 750 kb apart but more than 95% similar. The predicted genes *C38C10.3* and *F58A4.2*, mentioned above, are more than 200 kb apart but share a 1.4-kb region that is 98% similar. Two predicted exons are contained within the repeat. If the GENEFINDER predictions are correct, this would be an example of exon shuffling. Perhaps more likely, the *F58A4.2* version is a non-transcribed copy of part of a functional gene in *C38C10.3*. The 7.9-kb repeats in cosmid ZC262 described above contain segments of three different gene candidates, including a kinesin heavy-chain locus which was identified as *unc-116* (ref. 37). The *unc-116* gene begins outside the second copy of the repeat, with the last two exons present in the 7.9-kb. The same two exons in the first copy of the repeat are predicted to splice to a different exon in the second copy of the repeat. In addition to these recent gene duplications, we have identified similarities between other genes in the region. However, the degree of similarity suggests that these are more ancient in origin and may be examples of new gene families in *C. elegans*.

Prospects

The sequence reported here is already proving useful. In the very narrow sense, several genes previously under study have been sequenced, speeding the analysis and further study of these genes. The full sequence of the homeobox region will clarify its relationship to the *Drosophila Antennapedia* complex. More importantly, the 483 genes identified through genomic sequencing, together with the 1,194 genes discovered in the cDNA tagging project[27], are providing new and fruitful avenues for *C. elegans* research.

Furthermore, our experience in the pilot phase indicates that megabase-scale DNA sequencing at a reasonable cost is feasible with current methods and technology[6,48]. At the same time we feel that significant improvements are possible at almost every step. For example, we have developed an automated DNA template preparation capable of producing 400 M13 templates daily[18]. Initial plaque picking can also be done robotically[49], and instruments are under development that can perform large numbers of small-volume sequencing reactions automatically. Longer read lengths and greater sample capacity for the present generation of fluorescent gel readers are being developed, and more powerful instruments are being designed. Further improvements in the software will soon eliminate much of the sequence editing, which is currently a tedious task. Software tools are already available which simplify the selection of templates and oligonucleotides for directed sequencing (R.S., L.H. and S.D., unpublished). With these improvements and some increase in the scale of effort, production of more than 10 megabases of finished sequence per site per annum seems feasible. With this capacity in both halves of the consortium, the *C. elegans* genome sequence should be essentially completed before the end of 1998. In addition, both laboratories are contributing resources to speed the completion of the *S. cerevisiae* genome sequence. The complete genome sequences of these two organisms will provide insight into the genes that are likely to be common to all eukaryotes, and those specific to metazoans. □

Received 15 November 1993; accepted 5 January 1994.

1. Brenner, S. *Genetics* **77,** 71–94 (1974).
2. Wood, W. B. *et al. The Nematode* Caenorhabditis elegans (Cold Spring Harbor Laboratory Press, New York, 1988).
3. Coulson, A. R., Sulston, J. E., Brenner, S. & Karn, J. *Proc. natn. Acad. Sci. U.S.A.* **83,** 7821–7825 (1986).
4. Coulson, A., Waterston, R., Kliff, J., Sulston, J. & Kohara, Y. *Nature* **335,** 184–186 (1988).
5. Coulson, A. *et al. Bioessays* **13,** 413–417 (1991).
6. Sulston, J. *et al. Nature* **356,** 37–41 (1992).
7. Strauss, E. C., Kobori, J. A., Siu, G. & Hood, L. E. *Analyt. Biochem.* **154,** 353–360 (1986).
8. Deininger, P. *Analyt. Biochem.* **129,** 216–223 (1983).
9. Bankier, A. T. & Barrell, B. G. *Tech. Nucleic Acid Biochem.* B **5,** 1–34 (1983).
10. Smith, L. M. *et al. Nature* **321,** 674–679 (1986).
11. Connell, C. R. *et al. BioTechniques* **5,** 342–348 (1987).
12. Craxton, M. *Methods: A Comparison to Methods in Enzymology* Vol. 3 (ed. Roe, B.) 20–26 (Academic, San Diego, 1991).
13. Dear, S. & Staden, R. *Nucleic Acids Res.* **19,** 3907–3911 (1991).
14. Halloran, N., Du, Z. & Wilson, R. K. in *Methods in Molecular Biology* Vol. 10: *DNA Sequencing: Laboratory Protocols.* (eds Griffin, H. G. & Griffin, A. M.) 297–316 (Humana, Clifton, New Jersey, 1992).
15. Schriefer, L., Gebauer, B. K., Qiu, L. Q. Q., Waterson, R. H. & Wilson, R. K. *Nucleic Acids Res.* **18,** 7455–7456 (1990).
16. Lee, L. *et al. Nucleic Acids Res.* **20,** 2471–2483 (1992).
17. Hawkins, T. L., Du, Z., Halloran, N. D. & Wilson, R. K. *Electrophoresis* **13,** 552–559 (1992).
18. Watson, A., Smaldon, N., Lucke, R. & Hawkins, T. *Nature* **362,** 569–570 (1993).
19. Du, Z., Hood, L. & Wilson, R. K. *Meth. Enzym.* **218,** 104–121 (1993).
20. Gleeson, T. & Hillier, L. *Nucleic Acids Res.* **19,** 6481–6483 (1991).
21. Gleeson, T. J. & Staden, R. *Comput. appl. Biosci.* **7,** 398 (1991).
22. Hillier, L. & Green, P. *PCR Meth. Appl.* **1,** 124–128 (1991).
23. Dear, S. & Staden, R. *DNA Sequence* **3,** 107–110 (1992).
24. Alschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. *J. molec. Biol.* **215,** 403–410 (1990).
25. Koop, B. F., Rowan, L., Chen, W.-Q., Deshpande, P., Lee, H. & Hood, L. *BioTechniques* **14,** 442–447 (1993).
26. Krause, M. & Hirsh, D. *Cell* **49,** 753–761 (1987).
27. Waterston, R. *et al. Nature Genet.* **1,** 114–123 (1992).
28. McCombie, W. R. *et al. Nature Genet.* **1,** 124–131 (1992).
29. Green, P. *et al. Science* **259,** 1711–1716 (1993).
30. Burglin, T. R. *et al. Nature* **351,** 703 (1991).
31. Chisholm, A. *Development* **111,** 921–932 (1991).
32. Wang, B. B. *et al. Cell* **74,** 29–42 (1993).
33. Clark, S., Chisholm, A. & Horvitz, H. R. *Cell* **74,** 43–55 (1993).
34. Finney, M., Ruvkun, G. & Horvitz, H. R. *Cell* **55,** 757–769 (1988).
35. Guo, X., Johnson, J. J. & Kramer, J. M. *Nature* **349,** 707–709 (1991).
36. Naito, M., Kohara, Y. & Kurosawa, Y. *Nucleic Acids Res.* **20,** 2967–2969 (1992).
37. Patel, N., Thierry-Mieg, D. & Mancillas, J. R. *Proc. natn. Acad. Sci. U.S.A.* **90,** 9181–9185 (1993).
38. Weston, K., Yochem, J. & Greenwald, I. *Nucleic Acids Res.* **17,** 2138 (1989).
39. Yochem, J., Weston, K. & Greenwald, I. *Nature* **335,** 547–550 (1988).
40. Yochem, J. & Greenwald, I. *Cell* **58,** 553–563 (1989).
41. Waterston, R. H. GenBank locus CESUP5 (Acc. no. X54122) (1990).
42. Campbell, H. D. *et al. Proc. natn. Acad. Sci. U.S.A.* **90,** 11386–11390 (1993).
43. Waterston, R. H. & Brenner, S. *Nature* **275,** 715–719 (1978).
44. Sulston, J. E. & Brenner, S. *Genetics* **77,** 95–104 (1974).
45. Fichant, G. & Burks, C. *J. molec. Biol.* **220,** 659–671 (1991).
46. Collins, J., Forbes, E. & Anderson, P. *Genetics* **121,** 47–55 (1989).
47. Yuan, J., Finney, M., Tsung, N. & Horvitz, H. R. *Proc. natn. Acad. Sci. U.S.A.* **88,** 3334–3338 (1991).
48. Sulston, J. *Nature* **357,** 106 (1992).
49. Uber, D. C., Jaklevic, J. M., Theil, E. H., Lishanskaya, A. & McNeely, M. R. *BioTechniques* **11,** 642–647 (1991).

ACKNOWLEDGEMENTS. We thank T. Schedl for critical reading of the manuscript, P. Kassos and J. Rogers for administrative support, and other members of the *C. elegans* Genome Consortium for technical support. Database searches with St Louis sequences were done remotely using the NCBI BLAST server. ACEDB is available by anonymous ftp from cele.mrc-lmb.cam.ac.uk and from ncbi.nlm.nih.gov. This work was supported by the NIH National Center for Human Genome Research and the MRC Human Genome Mapping Project.